IMPROVED UNDERSTANDING OF THE RELATIVE QUALITY OF BUS PUBLIC TRANSIT USING A BALANCED APPROACH TO PERFORMANCE DATA NORMALIZATION

*Paper submitted for publication in the special edition of Transportation Research Part A: Policy and practice on new trends in analyzing and modeling public transport quality of service*

*Date submitted: 31st March 2017  
Date resubmitted: 31st October 2017*

**Corresponding author: Mark Trompet**  
Associate Director, Head of Bus and Airport Benchmarking  
Railway and Transport Strategy Centre, Centre for Transport Studies, Department of Civil and Environmental Engineering, Imperial College London  
Skempton Building, SW7 2AZ London; Tel +(20)75941519; Fax +(20)75946107  
E-mail: m.trompet@imperial.ac.uk

**Richard J. Anderson**  
Railway and Transport Strategy Centre, Centre for Transport Studies, Department of Civil and Environmental Engineering, Imperial College London  
Skempton Building, SW7 2AZ London; Tel +(20)75946088; Fax +(20)75946107  
E-mail: richard.anderson@imperial.ac.uk

**Daniel J. Graham**  
Railway and Transport Strategy Centre, Centre for Transport Studies, Department of Civil and Environmental Engineering, Imperial College London  
Skempton Building, SW7 2AZ London; Tel +(20)75946088; Fax +(20)75946107  
E-mail: d.j.graham@imperial.ac.uk

**Words: 6099**  
**Figures: 3**  
**Tables: 3**  
**Formulae: 5**
ABSTRACT

In order for bus operators and/or their respective authorities to understand where service quality can improve, it is useful to systematically compare performance with organizations displaying similarities in types of services offered, operational characteristics and density of the service area. These similar characteristics enable peer organizations to benchmark performance once their operational data are normalized for differences in scale of operations. The most commonly used normalization factors for the demand side output are passenger boardings and passenger kilometres. For the supply side output these are vehicle kilometres and vehicle hours. Through twelve years of experience in the International Bus Benchmarking Group (IBBG) a better understanding of differences in service characteristics between ‘similar’ peers has been achieved, which highlight a challenge for the interpretation of normalized performance. It became clear that relative performance should often not be concluded from performance indicators normalized in a single dimension. Variety between peers in commercial speed, trip length, vehicle planning capacity, vehicle weight and network efficiency result in the need for a bi-dimensional or balanced approach to data normalization. This paper quantifies the variety within these operational characteristics and provides examples of the interpretation bias this may lead to. A framework is provided for use by bus organization management, policymakers and benchmarking practitioners that suggests applicable combinations of denominators for a balanced normalization process, leading to improved understanding of relative performance.

KEYWORDS: Normalization, Benchmarking, Quality, Performance, Bus, Key Performance Indicators
1. INTRODUCTION

Based on a definition by Lema and Price (1) benchmarking is defined as a systematic process of continuously measuring, comparing and understanding organizations’ performance and change in performance of a diversity of key business processes against comparable peers anywhere else in the world to gain information which will help the participating organizations to take action to improve their performance.

One of the prerequisites for a successful benchmarking process is the selection of peer organizations based on similar characteristics. In bus benchmarking special focus should be given to the types of bus services offered (e.g. urban, suburban, charter, schoolbus, para-transit, etc), and the service area characteristics, in particular the density of operations and demand. These similarities enable organizations to compare performance once their data are normalized for scale.

As described by Trompet et al. (2), variability in comparable performance is welcomed as this results in the identification of best practices, which can lead to improved performance. However, the authors also describe that, even within a fixed group of seemingly comparable operators, there is variation in operating characteristics and environment that results in a subset of operators within that group being less directly comparable. The composition of this less directly comparable subset of operators will differ based on the key performance indicators (KPIs) produced. For example operators from low wage countries will perform well on financial efficiency KPIs when compared to peers operating in higher wage countries or cities, even when a Purchasing Power Parity index is applied. While trend analysis of all operators in the peer group will remain useful to identify best practices (e.g. cost reduction trends will pinpoint which operators are worth talking to for lessons on how to be more productive or efficient), the low wage cities primarily perform better in absolute terms than the high wage city operators due to the wage difference (which is primarily an exogenous factor) and are therefore less directly comparable in absolute financial performance achieved. Apart from differences in operating environments, a second important reason which could lead to peers in a single Key Performance Indicator to be less directly comparable is the bias due to the normalization factor chosen; this is the main focus of this research.

This paper provides evidence to bus organization management, policymakers and benchmarking practitioners that conclusions from relative quality and other performance differences observed in a single Key Performance Indicator (e.g. only normalized by a single denominator) can be biased. Due to variety between peers of operating characteristics such as commercial speed, trip length, vehicle capacity, vehicle weight and network efficiency (e.g. variety in amount of deadheading) it is necessary that performance is at least reviewed from two different dimensions, in order to obtain an improved and more balanced understanding of relative performance. This is especially important for bus organizations that have ‘extreme’ values in any of the five identified variable operating characteristics.

To achieve this objective the remainder of this paper is structured as follows. Section 2 reviews normalization factors used or suggested in a number of relevant previous and on-going bus public transport benchmarking initiatives. Section 3 describes the International Bus Benchmarking Group (IBBG) dataset which is used for this study. Section 4 describes the variability of five operational characteristics that can lead to a skewed perspective of relative performance when only one denominator is used for normalization. In section 5 a framework is then presented that schematically suggests which combinations of two denominators could be used to obtain a more realistic, balanced view of relative performance. The section also elaborates on
how this framework fits within the steps of the whole benchmarking process and how the resulting improved understanding of relative performance will help bus organization and/or authorities identify and subsequently prioritize areas for improvement and will facilitate the identification of peers which are worth reaching out to for information on best practices. Section 6 discusses the use of revenue service planning capacity kilometres as a normalization factor in more detail, as within the IBBG, this has especially resulted in improved performance comparability and understanding. Conclusions are drawn in Section 7.

2. NORMALIZATION IN TRANSPORT BENCHMARKING AND LITERATURE

An overview of public transport benchmarking initiatives has been provided in a variety of reports and papers (e.g. TCRP (3), EQUIP (4), Geerlings et al. (5) and Gudmundsson et al. (6)). With regards to urban bus performance measurement specifically, Mulley (7) describes the process and lessons learned from the UK Bus benchmarking Group which was based on the benchmarking handbook developed by EQUIP (8). Alongside the references mentioned above, the papers and project reports listed under references (9-21) have also been reviewed to understand how practitioners have normalized for scale. The focus in the review was to find recommendations towards improved normalization factors and on discussions in relation to the possible bias of using a single denominator for normalization.

The reports and papers generally underline the importance of peer selection and normalization for successful benchmarking. Complete sets (or examples) of main key performance indicators used are often provided. These confirmed that the most commonly used normalization factors for the demand side output are passenger boardings and passenger kilometres. For the supply side output these are vehicle kilometres and vehicle hours. Useful discussions have been provided on the use of vehicle hours as a preferred denominator over vehicle kilometres (Hencher and Daniels (9), Fielding et al. (10)). The majority of reviewed papers also mention the use of seat kilometres for supply side normalization. This led to the addition of Section 6 to this paper in which the authors discuss the role and possible bias of seat capacity kilometres as a normalization factor in a benchmarking exercise.

The benchmarking projects and literature reviewed did not explain in some level of detail why a certain normalization denominator was chosen (over an alternative one) and what bias could be involved in using that single denominator. This paper aims to contribute to this area.

Fielding et al. (11) and Stappenhorst (12) use cluster analysis to create comparable subgroups to improve comparability within performance indicators. This is feasible in a situation where the total number of participating organizations in a benchmarking group is sufficiently large; the size of each cluster of ‘more comparable’ peers should also be sufficient for performance comparison. The downside to clustering peers into more comparable subsets before benchmarking analysis is that possibly valid lessons and best practices from ‘lesser’ comparable peers are discarded. Clustering is done before the performance comparison is executed. This paper discusses a post performance comparison alternative, by understanding relative performance through a bi-dimensional normalization process.
3. THE DATA

The data used for this study have been collected through the International Bus Benchmarking Group (IBBG), which is facilitated by the Railway and Transport Strategy Centre at Imperial College London. The IBBG is now in its twelfth year. Its current members are TMB Barcelona, STIB Brussels, Dublin Bus, IETT Istanbul, Rapid Bus Kuala Lumpur, Carris Lisbon, London Buses, STM Montreal, NYCT & MTA Bus New York, RATP Paris, KCMT Seattle, STA Sydney Buses, Singapore SMRT and CMBC Vancouver. All agencies provide regular passenger public bus service operations in large urban areas. Up to fifteen years of data, from 2001 to 2015, are available for 105 key performance related data items to create KPIs in areas such as service availability, accessibility, reliability and quality, productivity, finance, safety and security, growth, learning and environmental performance. Some of these key performance related data items are broken down into further sub-categories such as vehicle type or outsourced versus in-house labor. These performance related data are supported by another 45 background data items to provide context and understanding. Only data from normal service operations is included, filtering out the effect of charter, tourist bus, para-transit and school bus services.

The focus of this research is on how the choice of a normalization factor impacts on the relative performance observed in a KPI. Within the IBBG data, there are 10 data items that are used as a denominator to normalize data for scale differences: passenger boardings, passenger kilometres, total vehicle kilometres, revenue vehicle kilometres, revenue planning capacity kilometres, total planning capacity kilometres, revenue vehicle hours, total vehicle hours, tonne kilometres and (categories of) staff hours. Each of these denominator data items are directly provided by the IBBG member operators as a single annual number, apart from total planning capacity kilometres, revenue planning capacity kilometres and tonne kilometres. The capacity kilometres denominators are calculated using the planning capacity data for all vehicle types in the fleet (as provided by the IBBG member operators) and the respective vehicle kilometres those vehicle types have run during the year, as shown in formulae 1 and 2 below. The tonne kilometres calculation uses the unloaded axel weight in tonnes of each vehicle type (as provided by the IBBG member operators) and the respective kilometres those vehicles types have run during the year, as shown in formula 3.

For each operator, the total planning capacity kilometres are calculated as follows:

$$\sum_{i=1}^{n} (ATVK_i \times SPC_i) \quad i = 1, ..., n$$  \hspace{1cm} (1)

For each operator, the revenue planning capacity kilometres are calculated as follows:

$$\left[ \sum_{i=1}^{n} (ATVK_i \times SPC_i) \right] \times \left( \frac{RVK}{TVK} \right) \quad i = 1, ..., n$$  \hspace{1cm} (2)

For each operator, the tonne kilometres are calculated as follows:

$$\sum_{i=1}^{n} (ATVK_i \times VT_i) \quad i = 1, ..., n$$  \hspace{1cm} (3)
Where
ATVK refers to annual total vehicle km, of the i-th vehicle type
SPC refers to service planning capacity, of the i-th vehicle type
VT refers to vehicle weight in tonne, of the i-th vehicle type
RVK refers to annual revenue vehicle km
TVK refers to annual total vehicle km

In the IBBG, it took three years of iterative definition development, data collection and analysis before the member operators were sufficiently satisfied with the level of comparability to be able to use the data for performance comparison. As data comparability is key in benchmarking, Trompet et al. (2) summarised the necessary benchmarking conditions and decisions in these first years of the IBBG to gradually increase the quality of the dataset. One of the main contributing factors to data quality is a strict confidentiality agreement which in combination with the willingness to help and learn from each other creates an open and honest information sharing environment. This confidentiality agreement also applied to the data used for this study. As a result, the graphs and tables in this paper have been anonymised and indexed where necessary. Nevertheless, the IBBG members agreed that the lessons from this study should be shared for the benefit of the wider bus public transport industry and policymakers.

4. VARIABILITY OF SERVICE CHARACTERISTICS

Through twelve years of experience in the IBBG it became clear that even between ‘similar’ peers differences in service characteristics exist. This is particularly relevant for service characteristics that are directly linked to the normalization factors used in bus performance benchmarking as described in the previous section. These five service characteristics are:

- Average passenger (unlinked) trip length, calculated as passenger kilometres divided by passenger boardings,
- Network efficiency, e.g. the proportion of deadheading kilometres/hours to total vehicle kilometres/hours. Frequently also expressed as the inverse, e.g. the proportion of revenue vehicle kilometres/hours to total vehicle kilometres/hours,
- Average weighted vehicle planning capacity of all buses in the fleet (see section 6 for a detailed discussion),
- Average commercial speed, calculated as actual revenue vehicle kilometres divided by actual revenue vehicle hours, and
- Average weighted vehicle weight, calculated as the weighted average unloaded axel weight in tonnes of all buses in the fleet

For these five service characteristics the extent of the variation between bus operators is determined to understand the potential effect a variable service characteristic can have on bus performance normalization. In Table 1, descriptive statistics are used to describe the variation of these five types of service characteristics using the 2015 data from the fifteen organizations in the IBBG database. These descriptive statistics are: Number of different bus operators in the sample (N), the mean value of the sample (µ), the minimum and maximum value within the sample and the standard deviation (σ). To be able to directly compare the extent of variability of the different service characteristics, the standard deviation is divided by the mean (σ / µ), to compute the
coefficient of variation (CV). The higher this number, the more variability is observed between values of bus organizations for that particular service characteristic.

The results in Table 1 are ranked by level of variability. The most variable service characteristic is average passenger trip length (CV=0.46), closely followed by network efficiency (CV=0.42). The range of values within these two characteristics is considerable. For example, the lowest average trip in the sample (2.8 km) is only 24% of the longest average passenger trip (11.6 km). As discussed in section 5, this has implications for performance benchmarking when data are only normalized for either passenger kilometres or passenger boardings.

**TABLE 1 Variability in Service Characteristics – 2015 data of the International Bus Benchmarking Group**

<table>
<thead>
<tr>
<th>Type of service characteristic</th>
<th>N</th>
<th>µ</th>
<th>Min</th>
<th>Max</th>
<th>σ</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average passenger trip length - km</td>
<td>15</td>
<td>5.3</td>
<td>2.8</td>
<td>11.6</td>
<td>2.5</td>
<td>0.46</td>
</tr>
<tr>
<td>Network efficiency - % of deadheading km</td>
<td>15</td>
<td>13.8</td>
<td>7.4</td>
<td>26.7</td>
<td>5.8</td>
<td>0.42</td>
</tr>
<tr>
<td>Average commercial speed – km/h</td>
<td>15</td>
<td>17.5</td>
<td>12.4</td>
<td>24.0</td>
<td>3.5</td>
<td>0.20</td>
</tr>
<tr>
<td>Weighted average vehicle planning capacity</td>
<td>15</td>
<td>67.6</td>
<td>53.1</td>
<td>92.2</td>
<td>11.9</td>
<td>0.18</td>
</tr>
<tr>
<td>Weighted average vehicle weight - tonne</td>
<td>15</td>
<td>13.2</td>
<td>11.2</td>
<td>17.0</td>
<td>1.6</td>
<td>0.12</td>
</tr>
</tbody>
</table>

N = Number of bus organizations in sample  
µ = Sample average  
Min = Minimum value  
Max = Maximum value  
σ = Standard deviation  
CV = Coefficient of variation

Two medium variable service characteristics are: commercial speed (CV=0.20) and service planning capacity (CV=0.18). Expressed in a coefficient of variation the average weighted vehicle weight is medium/low (CV=0.12), e.g. the majority of bus organizations operate a fleet of similar average weight. However the amplitude of data is significant, i.e. the minimum value is 11.2 tonne and the maximum value is 17.0 tonne. This gives indication that comparability of (in particular) fuel efficiency and emission KPIs could be improved by using weight as a normalization factor.

To understand if similar variety in service characteristics can be observed over time, Table 2 shows the coefficient of variation for these service characteristics for the period 2011-2015. For three operators vehicle hour data was available for 2014 and 2015 only; their data has therefore been removed regarding trends in the observed variety in commercial speed. The results in Table 2 clearly show that the variability level and differences observed in Table 1 can also be observed at very similar levels in other years.
When looking at the variability ranking, it seems that those service characteristics that are most variable are impacted by external factors, and factors that are harder to change or manage in the short or medium term. Trip length is determined by factors such as city size and density, availability of other transport modes and the complexity of the bus network itself. Network efficiency is strongly affected by depot locations and by the demand and supply profile. However, it is also affected by manageable impacts such as the decision to interline buses and/or the policy to bring buses back to the depot in the inter-peak period. Commercial speed is determined by factors such as traffic, dwell times and number of stops per route. Bus priority measures can be implemented on the short and medium term that speed up the organizations in the lower end of the spectrum, resulting in less variety in speed between organizations. Vehicle planning capacity is a function of the fleet characteristics and the passenger loading policy. Both can be managed on a medium term. Vehicle weights are less variable within the sample; however the significant range of average vehicle weights observed is strongly related to regional regulatory differences in vehicle specification, i.e. some bus organizations operate fleets that are considerably heavier than fleets used in other cities. Interesting observation is that variety in vehicle weight has been reduced annually from 2012 (CV=0.132) to 2015 (CV=0.117) as older heavier vehicles are replaced with newer lighter buses.

5. A BALANCED NORMALIZATION APPROACH

As described in section 3, the common scale denominators used in the IBBG to normalize data are: passenger boardings, passenger kilometres, total vehicle kilometres, revenue vehicle kilometres, revenue planning capacity kilometres, total planning capacity kilometres, revenue vehicle hours, total vehicle hours, tonne kilometres and (categories of) staff hours.

For each KPI, the most suitable denominator was chosen. For example, a number of labor productivity KPIs are normalized by total vehicle hours as hours are a more significant labor cost driver. Vehicle kilometres on the other hand are the main cost driver for maintenance activity and hence used as the preferred normalization factor.

Given the extent of variability of service characteristics, as identified in Table 1, within what is considered a group of comparable peers, it is important to realise that the choice of denominator can influence the relative performance (i.e. position in the graph) of individual bus organizations. If an organization has ‘extreme’ values in any of the five variable service

<table>
<thead>
<tr>
<th>Type of service characteristic</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average passenger trip length - km</td>
<td>15</td>
<td>0.427</td>
<td>0.432</td>
<td>0.450</td>
<td>0.470</td>
</tr>
<tr>
<td>Network efficiency - % of deadheading km</td>
<td>15</td>
<td>0.455</td>
<td>0.437</td>
<td>0.442</td>
<td>0.429</td>
</tr>
<tr>
<td>Average commercial speed – km/h</td>
<td>12</td>
<td>0.206</td>
<td>0.213</td>
<td>0.212</td>
<td>0.213</td>
</tr>
<tr>
<td>Weighted average vehicle planning capacity</td>
<td>15</td>
<td>NA</td>
<td>0.193</td>
<td>0.166</td>
<td>0.169</td>
</tr>
<tr>
<td>Weighted average vehicle weight - tonne</td>
<td>15</td>
<td>NA</td>
<td>0.132</td>
<td>0.128</td>
<td>0.124</td>
</tr>
</tbody>
</table>

N = Number of bus organizations in sample
characteristics mentioned in Table 1, the choice of denominator will affect the relative position to peers in the comparison.

To visualise this effect and to give an actual example of the possible bias with regards to interpreting relative bus transit performance Figure 1 shows for the 15 IBBG organizations in the sample, using 2015 data, the relative results for the same cost effectiveness KPI, but normalized by both passenger boardings and by passenger kilometres. Figure 1 is ranked by the indexed operating cost per passenger boarding, however the bus organizations are ranked alphabetically based on indexed operating cost per passenger kilometre (i.e. ‘B’ has the second highest cost per passenger kilometre). Where the ‘numerical’ rank does not match the ‘alphabetical’ rank this highlights which organisations show different relative performance in both KPIs. For five organisations (A, C, D, L and O) their relative performance is identical in both cost effectiveness KPIs. Their average trip lengths are average within this sample of organisations and therefore the choice of denominator is not creating a bias with regards to the interpretation of their relative performance to these peers. The other 10 organizations with extreme ‘trip length values, e.g. either very short or very long will perform differently in both cost effectiveness KPIs. Good examples are operator ‘M’ and ‘B’. When cost is normalised by passenger boarding these operators perform practically identical and have respectively the 6th and 7th highest cost per boarding within this group. However when normalised per passenger kilometre, operator B (with short average trip lengths) is now the 2nd most costly and operator M (with long average trip lengths) is now the 3rd most cost effective within this group, e.g. ranking 13th.

FIGURE 1 Operating Cost of IBBG Member Organizations normalized by Passenger Boardings vs Passenger Kilometres – Ranked by Operating Cost per Boarding
For completeness, Table 3 provides the results of similar analyses as described above and shown in Figure 1 for each of the five service characteristics identified in Table 1. The data from Figure 1 visualizes the normalization bias as a result of variability in trip length and this has been numerically represented as the first example in Table 3. For each of the other four service characteristics an example KPI was provided normalized by the two denominator data items that are linked to the respective service characteristic (see also Figure 2). Similar to the analysis in Figure 1 the bus organizations are ranked from left to right based on performance in KPI \(a\), and alphabetically (from A to O) based on performance in KPI \(b\). As in the case of Figure 1, all other four examples also show organisations for which the ‘positional’ rank does not match the ‘alphabetical’ rank highlighting which organisations show different relative performance in both KPIs.

It is important to state that the individual KPIs, as shown in Figure 1, or each of the examples in Table 3, remain valid in their own right; however each only answers part of a question, i.e. only ‘one side of the coin’ is shown. To understand the full picture and see ‘both sides of the coin’, two KPIs using complementary normalization factors (such as passenger boardings and passenger kilometres) need to be produced and simultaneously analyzed.

To assist bus organization management, policymakers and benchmarking practitioners a framework is provided in Figure 2 that suggests applicable combinations of denominators for a balanced normalization process, leading to improved understanding of relative performance.

FIGURE 2 Two Dimensional Normalization Framework for Bus Performance Data
TABLE 3 For Each of the Five Service Characteristics, One KPI Example Showing Relative Performance of Bus Organizations When Performance is Normalized by Either Factor Linked to the Respective Service Characteristic, 2015 Data of the International Bus Benchmarking Group

<table>
<thead>
<tr>
<th>Service Characteristic</th>
<th>Key Performance Indicator - 2015 indexed</th>
<th>Bus Organizations - Ranked from left to right based on performance in KPI a, and alphabetically (from A to O) based on performance in KPI b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trip Length</td>
<td>Operating Cost</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>1. Per Passenger Boarding</td>
<td>2.197</td>
</tr>
<tr>
<td></td>
<td>2. Per Passenger Kilometre</td>
<td>2.196</td>
</tr>
<tr>
<td>Network Efficiency</td>
<td>Operating Cost</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>1. Per Total Vehicle Hour</td>
<td>1.504</td>
</tr>
<tr>
<td></td>
<td>2. Per Revenue Vehicle Hour</td>
<td>1.580</td>
</tr>
<tr>
<td>Commercial Speed</td>
<td>Operating Cost</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>1. Per Revenue Vehicle Hour</td>
<td>1.580</td>
</tr>
<tr>
<td></td>
<td>2. Per Revenue Vehicle Km</td>
<td>1.181</td>
</tr>
<tr>
<td>Bus Planning Capacity</td>
<td>Vehicle Occupancy</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>1. Per Rev Planning Capacity Km</td>
<td>1.335</td>
</tr>
<tr>
<td></td>
<td>2. Per Revenue Vehicle Km</td>
<td>1.563</td>
</tr>
<tr>
<td>Vehicle Weight</td>
<td>Diesel Fuel</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>1. Per Total Diesel Vehicle Km</td>
<td>1.394</td>
</tr>
<tr>
<td></td>
<td>2. Per Diesel Vehicle Tonne Km</td>
<td>1.193</td>
</tr>
</tbody>
</table>
This framework shown in Figure 2 illustrates how different normalization factors, represented with black letters, interact with and complement each other. The five variable service characteristics identified in Table 1 are represented in the dark boxes with white letters. These service characteristics are placed on the border(s) of their relevant normalization factors. Every normalization factor is linked to a ‘counterpart’ normalization factor, which is needed to counterbalance the bias created by ‘extreme’ values in a service characteristic. This relation is shown through the arrows.

For example, an operator has a typical passenger trip length, vehicle weight and vehicle capacity. However they also have relatively high levels of dead running as a result of their depot locations and demand profile, and have relatively slow speeds compared to peers. Network efficiency and commercial speed are therefore their key areas that create an unbalanced perspective of their relative performance. Using the framework in Figure 2 this operator should therefore normalize performance data by both revenue vehicle kilometres and revenue vehicle hours to balance the effect of their relatively slow commercial speed. Similarly, they should reproduce any ‘revenue’ normalized KPI (kilometres, capacity kilometres or hours) also by its ‘total’ normalized counterpart (for differences in network efficiency).

The two white boxes stating ‘vehicle utilization performance’ and ‘system utilization performance’ are not essential parts of the framework itself, but do provide useful additional information on the relationship between passenger based normalization factors and vehicle based normalization factors.

In practical terms the bi-normalization framework should be used by bus operators, or other organizations that wish to benchmark their agency/city’s relative bus performance, within the third and 4th step of a complete benchmarking process. Generally speaking, the following steps are necessary for successful benchmarking:

1. Investigate/ensure that the database available is of sufficient quality for benchmarking purposes (as also addressed in the second last paragraph of section 7). Clean the database where necessary to ensure only comparable data remains.
2. Select peers from the remaining database which should be included in the performance comparison.
3. Create statistics for the five service characteristics shown in the bi-normalization framework and observe how the organisation is ranked amongst peers.
4. If the organization has any extreme values (compared to peers) in any of the five service characteristics, use the framework to follow the arrows from the respective ‘extreme’ characteristic(s) to understand which two performance normalization factors are linked.
5. Create KPI’s to measure the relative performance you wish to compare using the most applicable/intuitive normalization factor. However if this normalization factor is one that has been identified in step 4 as ‘extreme’, this KPI needs to be counterbalanced by creating a second KPI for the same performance measure, but normalized by the ‘counterpart’ normalization factor as shown in the bi-normalization framework.
6. Analyse the KPIs side by side as they each provide valuable information in their own right, but analysed together they provide the full picture in which the normalization bias is observed, but acknowledged and understood.

The resulting improved understanding of relative performance will help bus organizations and/or authorities identify and subsequently prioritize areas for improvement and will facilitate the identification of peers which are worth reaching out to for information on best practices.
It is important to note that labor hours, which are also used as a normalization factor for productivity and staff safety KPIs in the IBBG, have not been included in the normalization framework for balanced bus performance comparisons. Labor hours per employee can differ between organisations due to differences in contracts, benefits, local work rules and regulation and differences between bus operators in the use of overtime and part-time labor. For this reason it is recommended to use employee Full Time Equivalent (FTE) rather than employee headcount as the ‘per employee’ normalization factor. In the IBBG total labor hours are converted into standardized FTE’s by using one FTE is 2080 employee hours. Therefore the labor related normalization factors, e.g. ‘per labor hour’ and ‘per employee’ (as expressed in FTE) are directly related to each other and the performance comparison in either graph will show an identical result.

Apart from this methodological reason to not include employees / labor hours as denominators in the normalization model shown in Figure 2, it has also proven to be a challenge in the IBBG to obtain truly comparable labor hour data from bus operators. The main reason relates to the different levels of outsourcing of labor, and when outsourcing occurs, the varying levels of operator knowledge of the actual staff hours used by the contractors. Furthermore, there is the issue of reimbursed labor hours, e.g. driver hours for rail replacement service, maintenance labor hours to repair buses under warrantee, or accident repair labor which is paid for through an insurance company. Some operators do include these types of reimbursed labor hours in their annual data submissions, while others do not. Comparability of labor data is a current area of benchmarking research within the IBBG.

To confirm, all other common normalization factors mentioned in the first paragraph of this section are included in the framework shown in Figure 2.

6. VEHICLE PLANNING CAPACITY AS NORMALIZATION FACTOR FOR VEHICLE TYPE AND SIZE DIFFERENCES

The fourth most variable service characteristic identified in Table 1 is the average weighted service planning capacity of a vehicle. This section will discuss the improvements made with regards to understanding vehicle capacity and its use as a denominator.

The literature review, for example Hensher and Daniels (9), TCRP (14), MARETOPE (16) and Fielding et al. (19), confirmed that seat capacity kilometres is frequently used or proposed as a denominator to normalize for fleet capacity, especially for passenger loading key performance indicators. Some variations have been observed. Equip (8) uses the maximum capacity kilometres (seat and standing) in their load factor. It could not be confirmed if maximum here is defined as the manufacturers ‘licensed capacity’ or the physical ‘crush’ capacity. Phillips (13) and Fielding et al. (10) also mention total capacity kilometres, but the exact definition of ‘total’ is also not clear. The Urban Transport Benchmarking Initiative (18) defined capacity kilometres as place km, but only used those in the peak hour for a peak time loading KPI.

Harmonizing average weighted vehicle capacity in a benchmarking exercise is difficult as capacity can be defined differently. In the IBBG four types of vehicle capacity are collected annually: seating capacity, service planning capacity, licensed capacity and crush/max capacity. The difference between service planning and registered capacity is effectively the capacity that according to (service quality) policy should not be used. One can say it should not be considered to be available when scheduling bus services for service quality reasons.
Organizations have different policies on the proportion of registered capacity that can be used for service planning. For example, within the IBBG, one organization plans for their buses to be up to 80% full, but never more as they can be penalised by their authority. The service planning capacity is therefore 80% of the licensed capacity. Another organization, with longer average trip lengths, plans for most of their passengers to have a seat. Service planning capacity is therefore seating capacity + 30%, also called a load factor of 1.3. The Transit Capacity and Quality of Service Manual (22) describes that maximum scheduled passenger loads are typically 125 to 150% of seating capacity. Although sometimes inconsistent in policy, service planning capacity is always consistent in concept: e.g. the capacity offered to passengers. In the IBBG, after a pilot period, it was therefore decided that service planning capacity kilometres is the more suitable normalization factor for vehicle size and vehicle capacity. The remainder of this section provides the numerical evidence behind this decision.

Since ‘seat capacity kilometres’ is regularly proposed in literature as a normalization factor, it is useful to understand how seating capacity relates to service planning capacity. Figure 3 compares both the average weighted seating capacity per bus with the average weighted service planning capacity per bus across 15 IBBG bus organizations using the 2015 data. For each vehicle type in their fleet, organizations provided the number of buses in operable condition, the number of seats per bus and the planning capacity (seating + standees as defined by organization policy) per bus. This data was used to calculate the weighted average capacities per bus as shown in formula 4 and 5.

For each operator, the weighted average seating capacity is calculated as follows:

$$\frac{\sum_{i=1}^{n}(NB_i \times NS_i)}{TNB} \quad i = 1, \ldots, n$$  \hspace{0.5cm} (4)

For each operator, the weighted average service planning capacity is calculated as follows:

$$\frac{\sum_{i=1}^{n}(NB_i \times SPC_i)}{TNB} \quad i = 1, \ldots, n$$  \hspace{0.5cm} (5)

Where  
NB refers to number of buses, of the i-th vehicle type  
TNB refers to total number of buses  
SPC refers to service planning capacity per bus, of the i-th vehicle type  
NS refers to number of seats per bus, of the i-th vehicle type

In Figure 3, the organizations are ranked from left to right based on their weighted average seating capacity and ranked alphabetically based on their weighted average service planning capacity.

The first observation is that there is a substantial difference in the rank of most bus operators in both types of capacity. For example, operator ‘I’ has the second largest average weighted seating capacity per bus (54 seats), while ‘just’ the 9th largest service planning capacity per bus (allowing up to 61 people on their buses on average). Or operator ‘D’, which has the smallest average weighted seating capacity per bus (30 seats), and the 4th largest service planning capacity (allowing up to 75 people on their buses on average).

This variety observed in Figure 3 shows that in practice the average weighted seating capacity per bus is not the most effective normalization factor for vehicle size due to the different
type of vehicles used by operators. Double-decker and mini/midi buses have a high proportion of seats and a small amount of additional standing capacity, while standard and especially articulated buses have relatively few seats compared to the total capacity offered. Therefore by using seat kilometres as a denominator, a normalization discrepancy occurs due to fleet composition differences between bus operators.

![Average Weighted Vehicle Capacity - 2015](image)

**FIGURE 3** Average Weighted Vehicle Seating Capacity vs Average Weighted Vehicle Service Planning Capacity of the IBBG member organizations – ranked by seating capacity

Secondly, it can be concluded that the variety between the IBBG members within either type of capacity is quite significant. The average weighted seating capacity of operator ‘D’ is just 41% of the seating capacity of operator ‘A’. The variety within service planning capacity is significant (as shown in Table 1), but less extreme than the variety in seating capacity. The service planning capacity of operator ‘O’ is 58% of the service planning capacity of operator ‘A’. This makes the denominator more suitable for normalization as it reduces the potential bias in interpreting relative performance differences.

Overall the conclusion is that service planning capacity kilometres normalizes well for vehicle size, not only from an operational point of view, but also from a customer service/policy point of view.
7. CONCLUSIONS

The fifteen organizations of the International Bus benchmarking Group (IBBG) are all urban bus operators in large cities. They consider themselves comparable peers and successfully learn from each other through the exchange of best practices. Still, significant variety of operating characteristics has been observed, especially with respect to (in order of variability): passenger trip length, network efficiency, commercial speed, and vehicle capacity. While vehicle weights are less variable within the sample, a significant range has been observed strongly related to regulatory differences in vehicle specification.

The research shows that this variability in key service characteristics has implications for normalized performance comparisons, and therefore could lead to a biased interpretation of relative quality / performance. For example, passenger kilometres and passenger boardings are frequently used denominators to normalise for scale of demand. Since passenger kilometres are defined by the average trip length and the number of passenger boardings, variety of the average trip lengths between organizations becomes an important factor. The analysis showed that the shortest average trip length observed is just 24% of the longest average trip length in the sample. Therefore by only using either passenger kilometres or passenger boardings as a normalization factor, the relative performance of a bus operator observed in that single key performance indicator would be skewed if an operator has a particular low or high average trip length.

This paper illustrates that studying relative performance through a key performance indicator normalized by one normalization factor often only answers part of a question, i.e. it shows you ‘one side of a coin’. If your organization has average values in all four variable service characteristics mentioned above, the choice of denominator (for example vehicle kilometres or vehicle hours) will not affect the relative position to peers in the comparison, just which organizations are ranked on either side of you. However, if an organization has extreme values in any of the four more variable service characteristics identified, then the relative rank of that organization will significantly differ depending on the normalization factor used. To understand the full picture and see ‘both sides of the coin’, two KPIs using complementary normalization factors (such as vehicle kilometres and vehicle hours) need to be produced and simultaneously analyzed.

To assist bus organization management, policymakers and benchmarking practitioners a framework is provided that suggests applicable combinations of denominators for a balanced normalization process, leading to improved understanding of relative quality and other types of performance. This in turn would allow for better decision making with regards to allocation of resources and target setting.

Since vehicle capacity is an important normalization factor, section 6 of the paper explained the choice of the IBBG to use service planning capacity to normalize for vehicle size difference. The literature review confirmed that seat capacity kilometres is frequently used or proposed as a denominator to normalize for scale differences in vehicle size / capacity. The data from the IBBG, however, shows that in practice the average weighted seating capacity per bus is not the most effective normalization factor for vehicle size due to the different types of vehicles operated by operators and the different service planning policies in place. To overcome this, the IBBG uses vehicle service planning capacity as the preferred denominator; calculated as the weighted average number of seats per bus plus the additional weighted average number of standees that each operator plans for.
It is important to note that while the described bi-dimensional normalization framework for bus performance data can be used to improve the understanding of any bus performance comparison with any number of peers, the underlying quality and comparability of the data are naturally of the utmost importance. If the input data (either a numerator, or the normalization factor) is not comparable, the produced KPI using that data will not be comparable. Before understanding how the service characteristics of your organisation affects relative performance when normalized in only one dimension, attention must be given to understanding which data items in the available database are sufficiently comparable and can be used for analysis. The IBBG uses strict definitions for each data item collected and data is checked carefully and queried where necessary. However, there remain KPIs and data items for which comparability is challenging to achieve. For example in section 5 of this paper the data comparability issues observed with labor hour data are described. Based on experience in the IBBG, other areas that require further research into comparability of data items and Key Performance Indicators are on-time performance, safety/security and crowding indicators.

Finally, the bi-dimensional normalization framework for bus performance as presented in this paper can be adapted to be used for other transport modes. To do so, research is required to understand for each transport mode which service characteristics exist that directly impact normalization factors, and to understand the variability of these service characteristics. For example, in national/suburban rail, light rail/tram or metro benchmarking a service characteristic that varies between rail operators is the number of railcars per train as some operators run longer trains than others. Both train kilometres as well as car kilometres are frequently used normalization factors in rail benchmarking and therefore the (weighted) average number of cars per train will be an element to be added to a bi-dimensional normalization framework for rail performance.

ACKNOWLEDGEMENTS

The authors thank the members of the International Bus Benchmarking Group for their willingness to share the normalization improvements learned though the benchmarking group, with the wider transport industry.
REFERENCES


